CSC/DSC 265/465 Practice Problems - Midterms and Final Exams

CONTENTS:

Midterm - CSC 265 - March 22, 2016

NAME: _____

You are allowed two aid sheets on a standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). You have 1 hour and 10 minutes. Answer all five questions. All questions have equal weight. You are encouraged to read each question completely before starting.

1. We are given a multiple regression model, with sample size sample $n = 81$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

The following coefficient table is output:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3105	12.8428	0.569	0.571
x1	0.6285	0.9608	0.654	0.515
x2	0.8546	0.7705	1.109	0.271

However, an appropriate F-test shows that the full model significantly reduces the SSE compared to the null model $y = \beta_0 + \epsilon$. Suppose we are given the following error sums of squares SSE:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	88,748.85
M_1	$y = \beta_0 + \beta_1 x_1 + \epsilon$	82,952.39
M_2	$y = \beta_0 + \beta_2 x_2 + \epsilon$	82,112.42
M_{12}	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	81,664.43

Of the four models, which has the highest value of R^2, and which has the highest value of R^2_{adj}? Justify your answer in each case, and give the actual highest value.

[SOLN] The largest R^2 must be for the full model (ie it must have the smallest SSE). The total sum of squares $SSTO$ is the SSE for the null model M_0, so $SSTO = 88748.85$. So the largest R^2 is

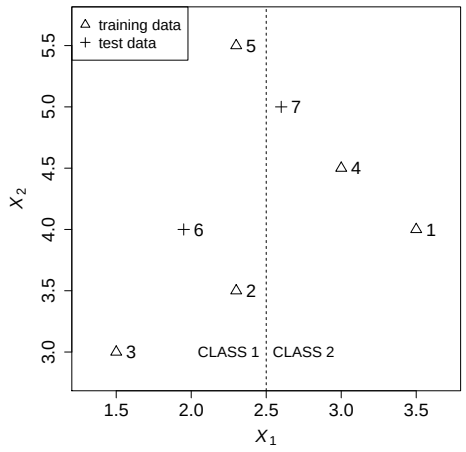
$$R^2[M_{12}] = 1 - \frac{SSE[M_{12}]}{SSTO} = 1 - \frac{81,664.43}{88,748.85} = 0.0798.$$

As for R^2_{adj}, this is zero for M_0. We can also eliminate M_1, since $SSE[M_1] > SSE[M_2]$, and both models have the same number of parameters. So,

$$\begin{aligned} R^2[M_2] &= 1 - \frac{SSE[M_2]/(n-2)}{SSTO/(n-1)} = 1 - \frac{82,112.42/79}{88,748.85/80} = 0.06307, \\ R^2[M_{12}] &= 1 - \frac{SSE[M_{12}]/(n-3)}{SSTO/(n-1)} = 1 - \frac{81,664.43/78}{88,748.85/80} = 0.05623. \end{aligned}$$

The largest value is $R^2[M_2] = 0.06307$.

2. To build a KNN classifier, the data in the following plot is used, partitioned into training and test data. That is, the test data is used to evaluate the accuracy of the KNN classifier built using the test data. As it happens, there are two classes, indicated in the plot by a class boundary. The pairwise distances are also given. By evaluating the classifier with the test data, estimate the classification errors for neighborhood sizes $K = 1$ and $K = 3$.



Pairwise Distance =

	1	2	3	4	5	6
2	1.300					
3	2.236	0.943				
4	0.707	1.221	2.121			
5	1.921	2.000	2.625	1.221		
6	1.550	0.610	1.097	1.163	1.540	
7	1.345	1.530	2.283	0.640	0.583	1.193

[SOLN] The correct classes for test observations $i = 6, 7$ are $y_i = 1, 2$.

For $K = 1$, observation $i = 6$, the neighborhood is $N = \{2\}$, so $\hat{y}_6 = 1$. For $i = 7$, $N = \{5\}$, $\hat{y}_7 = 1$. This means $CE = 0.5$.

For $K = 3$, observation $i = 6$, the neighborhood is $N = \{2, 3, 4\}$, so $\hat{y}_6 = 1$ (2/3 in N are class 1). For $i = 7$, $N = \{1, 4, 5\}$, $\hat{y}_7 = 2$ (2/3 in N are class 2). This means $CE = 0$.

3. We wish to develop a model which predicts the probability that a boxer wins a match based on weight differential. Suppose W is the amount in pounds by which the weight of boxer A exceeds the weight of boxer B (of course, W can be negative). Then

$$P(\text{Boxer } A \text{ wins} \mid W) = \frac{e^{\eta}}{1+e^{\eta}}, \text{ where } \eta = \beta_0 + \beta_1 W.$$

(a) Why would we expect

$$P(\text{Boxer } A \text{ wins} \mid W) + P(\text{Boxer } A \text{ wins} \mid -W) = 1$$

as a general rule? Show that this happens when $\beta_0 = 0$.

(b) Suppose we constrain $\beta_0 = 0$, and we are given $P(\text{Boxer } A \text{ wins} \mid 15) = 0.61$. What is β_1?

[SOLN]

(a) Suppose we have two boxers A, B. We expect $P(\text{Boxer } A \text{ wins}) + P(\text{Boxer } B \text{ wins}) = 1$. In addition, if $W = w$ for boxer A, then $W = -w$ for boxer B. Suppose $\beta_0 = 0$. Evaluate

$$\begin{aligned}
\frac{e^{\beta_1 w}}{1+e^{\beta_1 w}} + \frac{e^{-\beta_1 w}}{1+e^{-\beta_1 w}} &= \frac{e^{\beta_1 w}}{1+e^{\beta_1 w}} + \frac{e^{\beta_1 w}}{e^{\beta_1 w}} \times \frac{e^{-\beta_1 w}}{1+e^{-\beta_1 w}} \\
&= \frac{e^{\beta_1 w}}{1+e^{\beta_1 w}} + \frac{1}{1+e^{\beta_1 w}} \\
&= 1.
\end{aligned}$$

(b) We then have

$$P(\text{Boxer } A \text{ wins} \mid W) = \frac{1}{1+e^{-\beta_1 W}}.$$

If

$$0.61 = \frac{1}{1+e^{-\beta_1 15}},$$

then

$$\beta_1 = \frac{-1}{15} \log\left(\frac{1}{0.61} - 1\right) = 0.02982.$$

4. Suppose we may observe a vector of independent exponentially distributed random variables $X = (X_1, \ldots, X_m)$. Recall that the exponential density can be written

$$f(x) = \lambda e^{-\lambda x}, x > 0,$$

for any rate parameter $\lambda > 0$. The rate vector is then $\Lambda = (\lambda_1, \ldots, \lambda_m)$. Next, suppose we have a classification problem in which the vector of independent exponentially distributed random variables X comes from class A or B, defined by respective rate vectors $\Lambda_A = (\lambda_1^A, \ldots, \lambda_m^A)$ or $\Lambda_B = (\lambda_1^B, \ldots, \lambda_m^B)$.

(a) Suppose Λ_A, Λ_B are known. Suppose that the respective classes have prior probabilities π_A, π_B. Show that the Bayes classifier can be constructed, given observation $X = (X_1, \ldots, X_m)$, from two functions of X of the form:

$$h_A(X) = a_0 + \sum_{i=1}^{m} a_i X_i, \text{ and } h_B(X) = b_0 + \sum_{i=1}^{m} b_i X_i,$$

with the prediction being A if $h_A(X) > h_B(X)$ and B if $h_B(X) > h_A(X)$ (with the prediction made randomly when $h_B(X) = h_B(X)$). Express the coefficients a_i, b_i in terms of the parameters Λ_A, Λ_B and the prior probabilities π_A, π_B.

[SOLN] The joint distribution of $X = (X_1, \ldots, X_m)$ is

$$f(x_1, \ldots, x_m) = \prod_{i=1}^{m} \lambda_i e^{-\lambda_i x_i}.$$

The classifier function for class j may be given by

$$\begin{aligned} h_j(x_1, \ldots, x_m) &= = \log\left(f_j(x_1, \ldots, x_m)\pi_j\right) \\ &= \log(\pi_j) + \sum_{i=1} -\lambda_i x_i + \log(\lambda_i). \end{aligned}$$

So,

$$\begin{aligned} a_0 &= \log(\pi_A) + \sum_{i=1} \log(\lambda_i^A), \\ a_i &= -\lambda_i^A, \ i = 1, \ldots, m, \\ b_0 &= \log(\pi_B) + \sum_{i=1} \log(\lambda_i^B), \\ b_i &= -\lambda_i^B, \ i = 1, \ldots, m. \end{aligned}$$

5. We are given 2 classes, $j = 1, 2$. The distribution of a single dimensional observation is given by $X \sim N(\mu_j, \sigma_j^2)$, given classes $j = 1, 2$. Available estimates of μ_j are given by $\bar{X}_1 = 104.5$, $\bar{X}_2 = 56.7$. We assume $\sigma_1^2 = \sigma_2^2$, and a pooled estimate of the common variance is given by $s_{pooled}^2 = 7.82$. We accept as prior class probabilities $\pi_1 = 0.25, \pi_2 = 0.75$. Suppose an LDA classifier is constructed. Determine in which regions X predicts each class.

[SOLN] for LDA, the classifier is given by

$$\hat{y} = \text{argmax}_j h_j(x)$$

where

$$h_j(x) = x\mu_j/\sigma^2 - \frac{1}{2}\mu_j^2/\sigma^2 + \log(\pi_j).$$

The classification boundary x_b is the solution to $h_1(x_b) = h_2(x_b)$. There is only one, since the $h_j(x)$ are linear. This gives

$$x_b \times (104.5/7.82) - \frac{1}{2} \times 104.5^2/7.82 + \log(0.25) = x_b \times (56.7/7.82) - \frac{1}{2} \times 56.7^2/7.82 + \log(0.75)$$

or,

$$\begin{aligned} x_b \times \frac{104.5 - 56.7}{7.82} &= -\frac{1}{2} \times \frac{56.7^2 - 104.5^2}{7.82} + \log(0.75/0.25) \\ x_b \times 6.1125 &= 493.769, \end{aligned}$$

so that $x_b = 80.78$. Class $y = 1$ is predicted when $X > x_b$.

Midterm - CSC/DSC 265/465 - March 28, 2017

NAME: _____

You are allowed up to five aid sheets on standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). You have 1 hour and 10 minutes. Answer all five questions. All questions have equal weight. You are encouraged to read each question completely before starting.

Q1. We are given a multiple regression model, with sample size $n = 17$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Suppose we are given the following error sums of squares SSE for the model, and all reduced models:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	1,152,144.09
M_1	$y = \beta_0 + \beta_1 x_1 + \epsilon$	19,874.28
M_2	$y = \beta_0 + \beta_2 x_2 + \epsilon$	14,783.91
M_{1+2}	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	12,762.61
$M_{1\times 2}$	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$	12,291.50

Which model should be selected using the Bayesian information criterion (BIC)? Use the form $BIC = -2LL + C$, where LL is the log-likelihood function.

[SOLN] The BIC score is

$$BIC = -2LL + \log(n)d$$

where LL is the log-likleihood

$$LL = -\frac{n}{2}\log(SSE/n)$$

and d is the number of unknown parameters. If p is the number of model predictors (here, ranging from 0 to 3) it would be acceptable to set $d = p$, $p + 1$ or $p + 2$, the largest number incorporating parameters β_0, σ^2. However, any choice will yield exactly the same rankings. Here, we'll take $d = p$. This gives

$$BIC = n\log(SSE/n) + \log(n)p.$$

Using

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SSTO/(n-1)}$$

will also be accepted. Here, $SSTO$ equals SSE for the null model M_0. This gives calculations:

	Model	SSE	p	BIC	R^2_{adj}
1	M_0	1,152,144.09	1	189.1067	0.0000
2	M_1	19,874.28	2	122.9207	0.9816
3	M_2	14,783.91	2	117.8906	0.9863
4	M_{1+2}	12,762.61	3	118.2245	0.9873
5	$M_{1\times 2}$	12,291.50	4	120.4183	0.9869

Model M_2 is selected by the BIC score, while model M_{1+2} is selected by R^2_{adj}.

Q2. Under given conditions two sharp shooters $i = 1, 2$ are able to hit a target within a distance X, where X has an exponential density with mean μ_i (ie with density function $f(x) = \mu_i^{-1} \exp(-x/\mu_i)$, $x \geq 0$). Suppose we have independent observations of target accuracy $X_1, \ldots, X_n$ from one of the sharp shooters, and we wish to build a Bayesian classifier to predict that identity. Assume $\mu_1 < \mu_2$ are known, and that the prior probabilities are π_1, π_2.

Show that the classifier requires only the sum $S = \sum_{i=1}^{n} X_i$, and give precise conditions under which indentity $i = 1$ would be predicted.

[SOLN] For each class $i = 1, 2$ the joint distribution of $X = (X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} \mu_i^{-1} e^{-x_k/\mu_i}.$$

The classifier function for class i may be given by

$$\begin{aligned} h_i(x_1, \ldots, x_n) &= \log\left(f_i(x_1, \ldots, x_n)\pi_i\right) \\ &= \log(\pi_i) - n\log(\mu_i) - \sum_{k=1}^{n} x_k/\mu_i. \end{aligned}$$

Class $i = 1$ is predicted if $h_1(x_1, \ldots, x_n) - h_2(x_1, \ldots, x_n) > 0$, which is equivalent to

$$\left[\log(\pi_1) - n\log(\mu_1) - \sum_{k=1}^{n} x_k/\mu_1\right] - \left[\log(\pi_2) - n\log(\mu_2) - \sum_{k=1}^{n} x_k/\mu_2\right] > 0,$$

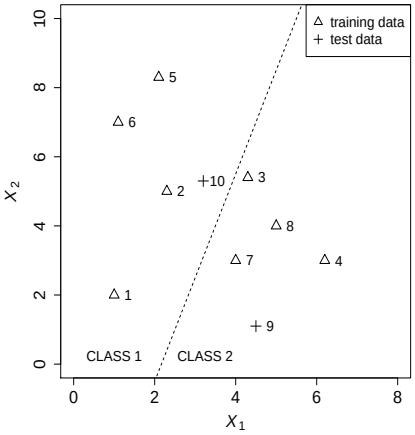
or, more concisely,

$$\sum_{k=1}^{n} x_k < \frac{\log(\pi_1/\pi_2) - n\log(\mu_1/\mu_2)}{\mu_1^{-1} - \mu_2^{-1}}$$

Q3. To build a KNN classifier, the data in the following plot is used, partitioned into training and test data. That is, the test data is used to evaluate the accuracy of the KNN classifier built using the training data. As it happens, there are two classes, indicated in the plot by a class boundary. The pairwise distances are also given. Give the predicted class for each test observation (labelled 9 and 10), using neighborhood sizes $K = 1$ and $K = 3$. Show clearly how these were obtained.

Pairwise Distance =

	1	2	3	4	5	6	7	8	9	10
1	0.00	3.27	4.74	5.30	6.40	5.00	3.16	4.47	3.61	3.97
2	3.27	0.00	2.04	4.38	3.31	2.33	2.62	2.88	4.48	0.95
3	4.74	2.04	0.00	3.06	3.64	3.58	2.42	1.57	4.30	1.10
4	5.30	4.38	3.06	0.00	6.70	6.48	2.20	1.56	2.55	3.78
5	6.40	3.31	3.64	6.70	0.00	1.64	5.63	5.19	7.59	3.20
6	5.00	2.33	3.58	6.48	1.64	0.00	4.94	4.92	6.81	2.70
7	3.16	2.62	2.42	2.20	5.63	4.94	0.00	1.41	1.96	2.44
8	4.47	2.88	1.57	1.56	5.19	4.92	1.41	0.00	2.94	2.22
9	3.61	4.48	4.30	2.55	7.59	6.81	1.96	2.94	0.00	4.40
10	3.97	0.95	1.10	3.78	3.20	2.70	2.44	2.22	4.40	0.00



[SOLN] The correct classes for test observations $i = 9, 10$ are $y_i = 2, 1$, respectively.

For $K = 1$, observation $i = 9$, the neighborhood is $N = \{7\}$, so $\hat{y}_9 = 2$ [correct], since observation 7 is class 2. For $i = 10$, $N = \{2\}$, $\hat{y}_{10} = 1$ [correct], since observation 2 is class 1.

For $K = 3$, observation $i = 9$, the neighborhood is $N = \{4, 7, 8\}$, so $\hat{y}_9 = 2$ [correct], since all observations in N are class 2. For $i = 10$, $N = \{2, 3, 8\}$, $\hat{y}_{10} = 2$ [incorrect], since 2/3 in N are class 2.

Q4. A model for predicting victory for a professional baseball team is developed. It depends on two predictors, $T =$ temperature on game day, in degrees Fahrenheit, and the indicator variable $I_H = 1$ for home games. The following logistic regression model is used:

$$P(\text{Win Game}) = \frac{e^{\eta}}{1+e^{\eta}}, \text{ where } \eta = \beta_0 + \beta_1 T + \beta_2 I_H + \beta_3 T \times I_H.$$

Suppose the parameter estimates are $\hat{\beta}_0 = -0.025$, $\hat{\beta}_1 = -0.027$, $\hat{\beta}_2 = -2.15$, $\hat{\beta}_3 = 0.076$.

(a) When playing at home, does the team prefer higher or lower temperatures? What about when they play away?

(b) Suppose the temperature is assumed to be within the range $[65, 85]$. What are the minimum and maximum values of $P(\text{Win Game})$ when the team plays at home, and when the team plays away?

(c) In order to predict the number of wins for a season, a simple temperature model is developed. The temperature will be either 65° or 85°. We assume probabilities $P(T = 85^\circ \mid I_H = 1) = 0.45$, $P(T = 85^\circ \mid I_H = 0) = 0.62$. Assuming exact half of the games are home games, what is the overall predicted win rate for the season?

[SOLN]

(a) First, note that $P(\text{Win Game})$ is an increasing function of η. When playing at home, $I_H = 1$, so

$$\eta = \beta_0 + \beta_2 + (\beta_1 + \beta_3)T.$$

We have estimate $\beta_1 + \beta_3 \approx \hat{\beta}_1 + \hat{\beta}_3 = -0.027 + 0.076 = 0.049$. In this case, η, and therefore $P(\text{Win Game})$, increases with temperature T. On the other hand, for away games $I_H = 0$, so

$$\eta = \beta_0 + \beta_1 T.$$

Since $\beta_1 \approx \hat{\beta}_1 = -0.027$, we conclude that η, and therefore $P(\text{Win Game})$, decreases with temperature T.

(b) In general, we have

$$\begin{aligned} P(\text{Win Game} \mid I_H = 0) &= (1 + \exp(0.025 + 0.027 \times T))^{-1} \\ P(\text{Win Game} \mid I_H = 1) &= (1 + \exp(2.175 - 0.049 \times T))^{-1} \end{aligned}$$

For both home and away games, the extreme points are calculated at $T = 65, 85$. We need the probabilities

$$\begin{aligned} P(\text{Win Game} \mid T = 65, I_H = 0) &= (1 + \exp(0.025 + 0.027 \times 65))^{-1} = 0.1443 \\ P(\text{Win Game} \mid T = 85, I_H = 0) &= (1 + \exp(0.025 + 0.027 \times 85))^{-1} = 0.0895 \\ P(\text{Win Game} \mid T = 65, I_H = 1) &= (1 + \exp(2.175 - 0.049 \times 65))^{-1} = 0.7330 \\ P(\text{Win Game} \mid T = 85, I_H = 1) &= (1 + \exp(2.175 - 0.049 \times 85))^{-1} = 0.8797. \end{aligned}$$

(c) The overall probability is found by the law of total probability:

$$\begin{aligned} p_W &= P(\text{Win Game} \mid T = 65, I_H = 0)P(T = 65, I_H = 0) + P(\text{Win Game} \mid T = 85, I_H = 0)P(T = 85, I_H = 0) + \\ &\quad P(\text{Win Game} \mid T = 65, I_H = 1)P(T = 65, I_H = 1) + P(\text{Win Game} \mid T = 85, I_H = 1)P(T = 85, I_H = 1) \\ &= [0.1443 \times (1 - 0.62) + 0.0895 \times 0.62 + 0.7330 \times (1 - 0.45) + 0.8797 \times 0.45] / 2 \\ &\approx 0.454. \end{aligned}$$

Q5. We are given 2 classes, $j = 1, 2$. The distribution of a single dimensional observation is given by $X \sim N(\mu_j, \sigma_j^2)$, given classes $j = 1, 2$. Available estimates of μ_j are given by $\bar{X}_1 = 44.5$, $\bar{X}_2 = 20.7$. We assume $\sigma_1^2 = \sigma_2^2$, and a pooled estimate of the common variance is given by $s^2_{pooled} = 3.82$. We accept as prior class probabilities $\pi_1 = 0.4, \pi_2 = 0.6$. Suppose an LDA classifier is constructed. Determine in which regions for which X predicts each class.

[SOLN] For LDA, the classifier is given by

$$\hat{y} = \text{argmax}_j h_j(x)$$

where

$$h_j(x) = x\mu_j/\sigma^2 - \frac{1}{2}\mu_j^2/\sigma^2 + \log(\pi_j).$$

The classification boundary x_b is the solution to $h_1(x_b) = h_2(x_b)$. There is only one, since the $h_j(x)$ are linear. This gives, after substituting the estimates,

$$x_b \times (44.5/3.82) - \frac{1}{2} \times 44.5^2/3.82 + \log(0.4) = x_b \times (20.7/3.82) - \frac{1}{2} \times 20.7^2/3.82 + \log(0.6)$$

or,

$$\begin{aligned} x_b \times \frac{44.5 - 20.7}{3.82} &= -\frac{1}{2} \times \frac{44.5^2 - 20.7^2}{3.82} + \log(0.6/0.4), \\ x_b &= 32.665, \end{aligned}$$

so that class $y = 1$ is predicted when $X > x_b = 32.665$.

Midterm Practice Questions - CSC 265 - Spring 2016

1. A risk score for consumer bankruptcy is developed. The range is $x \in [0, 100]$, with $x = 100$ representing the highest risk. To calibrate the score a logistic regression model is fit using observed pairs (y_i, x_i), where $y_i = 1$ denotes bankruptcy observed within a year, and x is the risk score at the beginning of the observed year. The linear prediction term is $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0 = -13.35$, $\hat{\beta}_1 = 0.17$.

   (a) At which score x is the yearly probability of bankruptcy equal to 50%?

   (b) What is the yearly probability of bankruptcy for the maximum score $x = 100$?

   (c) Suppose $(0.14, 0.20)$ is a 95% confidence interval for $\hat{\beta}_1$. Give a 95% confidence interval for the odds ratio for yearly bankruptcy between consumers with risk scores of $x = 70$ and $x = 40$.

2. A KNN classifier uses as training data $n_1 = 35$, $n_2 = 56$, $n_3 = 97$ observations from classes 1,2 and 3. Suppose the neighbourhood size K is allowed to vary from 1 to 187 (the total sample size - 1). LOO cross validation is used to estimate the classification error CE for each K. To what value does CE approach as K approaches 187?

3. Given a single predictor x and response y, a polynomial regression model is considered:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon.$$

   Suppose we consider four models (full model and 3 reduced models). Suppose further that the sample size is $n = 41$, and that the four models are fit, yielding the following error sums of squares SSE:

Table 1:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	229.26
M_1	$y = \beta_0 + \beta_1 x + \epsilon$	35.388
M_2	$y = \beta_0 + \beta_2 x^2 + \epsilon$	31.923
M_{12}	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$	31.294

   Suppose we rank the four models according to the following two rules:

   Rule 1 If two models M and M' have the same number of predictors, the one with the smaller SSE is ranked higher.

   Rule 2 If model M' can be obtained from model M by adding a single predictor, and the p-value for an F-test comparing full model M' to reduced model M is ≤ 0.05, then M' is ranked higher than M, otherwise it is ranked lower.

   The selected model is the one of highest rank. Assuming the models of Table 1 can be consistently ranked, what is the highest ranked model?

4. Consider the following classification problem. An object can be classified as Red, Green or Blue. Instead of observing the class, we observe an index $I \in \{1, 2, 3, 4\}$. The conditional probabilities of I given each class are given in Table 2.

Table 2:

$I =$	1	2	3	4
Red	0.6	0.2	0.1	0.1
Green	0.1	0.9	0	0
Blue	0.1	0.1	0.6	0.2

   Give the Bayes classifier based on observed index I for the following two cases.

(a) The class prior probabilities are $\pi_{Red} = \pi_{Green} = \pi_{Blue} = 1/3$.

(b) The class prior probabilities are $\pi_{Red} = 0.5$, $\pi_{Green} = 0.4$, $\pi_{Blue} = 0.1$.

5. We are given 2 classes, $j = 1, 2$. The distribution of a single dimensional observation is given by $X \sim N(\mu_j, \sigma_j^2)$, given classes $j = 1, 2$. Available estimates of μ_j, σ_j^2 are $\bar{X}_1 = 34.5$, $\bar{X}_2 = 43.7$, $s_1^2 = 3.45$, $s_2^2 = 7.56$, based on samples of size $n_1 = 54$, $n_2 = 34$.

   (a) Suppose we construct an LDA classifier based on X. Determine in which regions X predicts each class.

   (b) Suppose we construct an QDA classifier based on X. Determine in which regions X predicts each class.

6. Suppose we have model

$$y = f(x) + \epsilon$$

   where $E[\epsilon] = 0$. We are given the following 5 paired observations of (x_i, y_i) sampled from this model (Table 3).

Table 3:

i	x_i	y_i
1	3	5
2	6	4
3	9	3
4	12	8
5	15	7

   Construct and sketch a KNN regression estimate of $f(x)$ on the range $[3, 15]$ using a neighborhood size of $K = 3$.

Midterm Practice Questions [Solutions] - CSC 265 - Spring 2016

1. The prediction takes the form:

$$P_x(y=1) = \frac{1}{1+\exp(-\hat{\beta}_0 - \hat{\beta}_1 x)} = \frac{1}{1+\exp(13.35 - 0.17x)}.$$

   (a) $P_x(y=1) = 0.5$ when $13.35 - 0.17x = 0$, or $x = 13.35/0.17 = 78.53$.

   (b) We have

$$P_{100}(y=1) = \frac{1}{1+\exp(13.35 - 0.17 \times 100)} = 0.975.$$

   (c) We have

$$OR = e^{\beta_1(70-40)} = e^{\beta_1 30}.$$

   If $(0.14, 0.20)$ is a 95% confidence interval for β_1, the equivalent confidence interval for OR is

$$CI = \left(e^{0.14\times 30}, e^{0.20\times 30}\right) = (66.69, 403.43).$$

2. When K is large enough, all observations in the training sample will be in the $K-$neighbourhood, so the prediction will always be $\hat{y} = 3$, since $y = 3$ is the highest frequency class. Of the LOOCV predictions, 97/188 will correctly predict class 3, the remaining will incorrectly predict class 3. So $CE = 91/188 \approx 48.4\%$.

3. Write $M' > M$ if M' is ranked higher than M.

Table 1:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	229.26
M_1	$y = \beta_0 + \beta_1 x + \epsilon$	35.388
M_2	$y = \beta_0 + \beta_2 x^2 + \epsilon$	31.923
M_{12}	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$	31.294

   (a) Directly from Table 1 we have $M_2 > M_1$, since M_2 has the smaller SSE.

   (b) The F-statistic for comparing M_2 and M_0 is

$$F = \frac{(SSE_0 - SSE_2)}{SSE_2/(n-2)} = \frac{(229.26 - 31.923)}{31.923/39} = 241.0846$$

   The appropriate critical value is $F_{1,39;\alpha} = 4.091$, so we conclude $M_2 > M_0$.

   (c) The F-statistic for comparing M_{12} and M_2 is

$$F = \frac{(SSE_2 - SSE_{12})}{SSE_{12}/(n-3)} = \frac{(31.923 - 31.294)}{31.294/38} = 0.764$$

   The appropriate critical value is $F_{1,38;\alpha} = 4.098$, so we conclude $M_2 > M_{12}$.

   This analysis suffices to conclude that M_2 is the highest ranked model.

4. Given that we observe random index $I = i$, the Bayes classifier is

$$\hat{y} = \text{argmax}_{y\in\{Red,Green,Blue\}} P(I = i \mid y)\pi_y$$

   where, for example $P(I = 2 \mid Green)\pi_{Green} = 0.9 \times \pi_{Green}$.

(a) To construct the Bayes classifier for a uniform prior, we only need to select the highest class probability for each outcome of I (Table 2), giving

$$\begin{aligned}\hat{y}(1) &= \textit{Red}\\ \hat{y}(2) &= \textit{Green}\\ \hat{y}(3) &= \textit{Blue}\\ \hat{y}(4) &= \textit{Blue}.\end{aligned}$$

Table 2:

$I=$	1	2	3	4
Red	0.6	0.2	0.1	0.1
Green	0.1	0.9	0	0
Blue	0.1	0.1	0.6	0.2

(b) We can construct the Bayes classifier for any prior probabilities in the same way, by multiplying each entry in Table 2 by the respective class probability. See Table 3.

Table 3:

$I=$	1	2	3	4
Red	0.3	0.1	0.05	0.05
Green	0.04	0.36	0	0
Blue	0.01	0.01	0.06	0.02

This gives Bayes classifier

$$\begin{aligned}\hat{y}(1) &= \textit{Red}\\ \hat{y}(2) &= \textit{Green}\\ \hat{y}(3) &= \textit{Blue}\\ \hat{y}(4) &= \textit{Red}.\end{aligned}$$

5. (a) For LDA, the classifier is given by

$$\hat{y} = \text{argmax}_j h_j(x)$$

where

$$h_j(x) = x\mu_j/\sigma^2 - \frac{1}{2}\mu_j^2/\sigma^2 + \log(\pi_j).$$

To estimate the common variance σ^2 we use pooled estimate

$$s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{53\times 3.45 + 34\times 7.56}{86} = 5.027.$$

To estimate the prior probabilities we may use observed frequencies $\hat{\pi}_1 = 54/88$, $\hat{\pi}_2 = 34/88$. To estimate μ_j, we use the reported sample means. The classification boundary x_b is the solution to $h_1(x_b) = h_2(x_b)$. There is only one, since the $h_j(x)$ are linear. This gives

$$x_b \times (34.5/5.027) - \frac{1}{2}34.5^2/5.027 + \log(54/88) = x_b \times (43.7/5.027) - \frac{1}{2}43.7^2/5.027 + \log(34/88),$$

So that $x_b = 39.35$. Class $y = 2$ is predicted when $X > x_b$.

(b) For QDA, the classifier is given by

$$\hat{y} = \text{argmax}_j h_j(x)$$

where

$$h_j(x) = -\frac{1}{2}(x-\mu_j)^2/\sigma_j^2 - \frac{1}{2}\log(\sigma_j^2) + \log(\pi_j).$$

We use the same estimates as for the LDA classifier, except that the σ_j^2 are estimated separately by s_j^2, for $j = 1, 2$. Then $y = 2$ is predicted when

$$h_2(x) - h_1(x) > 0,$$

which is a quadratic equation $h_2(x) - h_1(x) = ax^2 + bx + c$, where

$$\begin{aligned} a &= -\frac{1}{2}\left[\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right], \\ b &= \left[\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right], \\ c &= -\frac{1}{2}\left[\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2}\right] - \frac{1}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \log\left(\frac{\pi_2}{\pi_1}\right). \end{aligned}$$

Substituting the estimates we get $a = 0.0788, b = -4.22, c = 45.34$. The boundary equation $h_2(x) - h_1(x)$ has two solutions, $x_b = 14.88, 38.67$. Since $a > 0$, we have $h_2(X) - h_1(X) > 0$ when $X < 14.88$ or $X > 38.67$, which forms the prediction region for class $y = 2$.

6. To construct $\hat{f}(x)$, for any x identify the 3 values from $\{3, 6, 9, 12, 15\}$ nearest x, then $\hat{f}(x)$ is the average response y_i paired with the three values x_i. Then $\hat{f}(x)$ is a step function which may only have discontinuities at $x = 7.5$ and $x = 10.5$. The resulting fitted function is shown in Figure 1 (discontinuities are indicated by vertical lines).

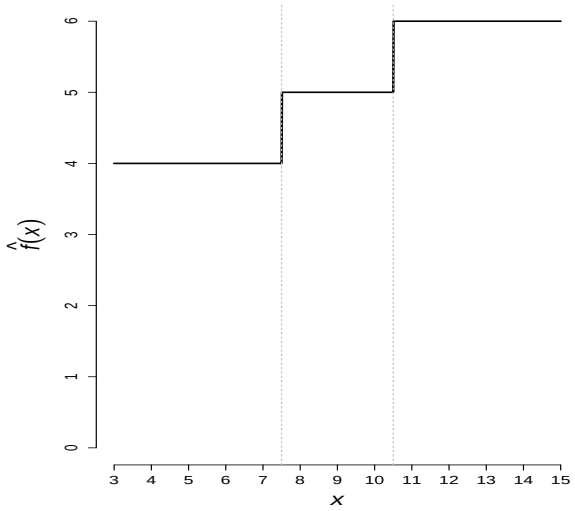


Figure 1: Problem 6

Final Exam - CSC 265 - May 6, 2016

NAME: _____

You are allowed two aid sheets on standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). You have 3 hours. Answer all ten questions. All questions have equal weight. You are encouraged to read each question completely before starting.

1. The following full linear regression model is considered:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

An all subsets model selection procedure is to be used to determine which of the predictors X_1, X_2, X_3 to retain. The sample size is $n = 150$. The AIC score will be used, in the form $AIC = n \log(SSE/n) + C$. The error sum of squares for each tested model is given in the following table. What will be the selected model?

	Model	SSE
1	$Y = 1$	1940.373
2	$Y = X_1$	1399.735
3	$Y = X_2$	485.242
4	$Y = X_3$	1940.367
5	$Y = X_1 + X_2$	0.268
6	$Y = X_1 + X_3$	1399.713
7	$Y = X_2 + X_3$	485.239
8	$Y = X_1 + X_2 + X_3$	0.267

SOLUTION: Formula is

$$AIC = n \log(SSE/n) + 2k,$$

where k is the number of parameters. We can construct table:

	Model	k	SSE	AIC
1	$Y = 1$	1	1940.37	386.00
2	$Y = X_1$	2	1399.74	339.01
3	$Y = X_2$	2	485.24	180.10
4	$Y = X_3$	2	1940.37	388.00
5	$Y = X_1 + X_2$	3	0.27	-943.35
6	$Y = X_1 + X_3$	3	1399.71	341.01
7	$Y = X_2 + X_3$	3	485.24	182.10
8	$Y = X_1 + X_2 + X_3$	4	0.27	-941.41

The model $Y = X_1 + X_2$ has the lowest AIC (= -943.35).

2. A regression model is to be developed for a response Y and single predictor X in range $X \in [100, 200]$, based on $n = 11$ paired observations. The following two models were considered, and the resulting error sum of squares is reported:

   (a) Simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ [SSE = 67813.35].

   (b) Cubic spline with single knot at $X = 150$ [SSE = 53599.91].

   Using the BIC score, which is the best model? Use the form $BIC = n \log(SSE/n) + C$.

   SOLUTION: Formula is

   $$BIC = n \log(SSE/n) + \log(n)k,$$

   where k is the number of parameters. We can construct table:

Model	k	SSE	BIC
linear	2.00	67813.35	100.79
cubic spline	5.00	53599.91	105.39

   The linear model has the lowest BIC (= 100.79).

3. Two variables Y and X are believed to have the following relationship:

$$Y = aX^b$$

for two constants a, b. According to a certain conjecture, Y is proportional to the square root of X. In order to resolve this question paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled, where $n = 51$. The simple linear regression model

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

is fit, with the following output:

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.3137646 0.02512828 92.078108 1.447613e-56
log(x)      0.4985705 0.05774591  8.633867 2.084971e-11
```

Formulate appropriate null and alternative hypotheses for this question in terms of the regression coefficients β_0 and/or β_1. Is there evidence at an $\alpha = 0.05$ significance level with which to reject the conjecture? [Use $t_{49,0.025} = 2.01$]

SOLUTION: The hypotheses are

$$H_o : \beta_1 = 1/2 \text{ against } H_a : \beta_1 \neq 1/2.$$

The appropriate t-statistic is

$$T = \frac{\hat{\beta}_1 - 1/2}{SE_{\hat{\beta}_1}} = \frac{0.4985705 - 1/2}{0.05774591} \approx -0.0247.$$

Since $|T| < t_{49,0.025}$ we do not reject the conjecture at significance level $\alpha = 0.05$.

4. Suppose we have $n = 4$ observations of features, for which the following distance matrix is calculated:

	1	2	3	4
1	0	122	76	53
2	122	0	46	69
3	76	46	0	23
4	53	69	23	0

Using the average distance agglomeration method, construct a hierarchical cluster for this data. Sketch a dendogram, indicating precisely the height of each node.

SOLUTION:

The compact distance between two clusters A and B is

$$D(A, B) = \frac{1}{|A||B|} \sum_{i \in A, j \in B} d_{ij}.$$

To construct the clustering, we use the following steps:

(a) Start with clusters {1},{2},{3},{4}.

(b) First join the two nearest observations, which are 3 and 4 ($d_{3,4} = 23$). This gives clusters {1}, {2}, and $\{3, 4\}$ joined at distance 23.

(c) The cluster distances are now

$$\begin{aligned} D(\{1\}, \{2\}) &= d_{1,2} = 122, \\ D(\{1\}, \{3, 4\}) &= (d_{1,3} + d_{1,4})/2 = (76 + 53)/2 = 64.5, \\ D(\{2\}, \{3, 4\}) &= (d_{2,3} + d_{2,4})/2 = (46 + 69)/2 = 57.5. \end{aligned}$$

The smallest cluster distance is $D(\{2\}, \{3, 4\}) = 57.5$, so combine clusters $\{2\}$ and $\{3, 4\}$. This gives clusters {1}, and $\{2, 3, 4\}$, joined at distance 57.5.

(d) The remaining clusters are joined. They have cluster distance

$$D(\{1\}, \{2, 3, 4\}) = (d_{1,2} + d_{1,3} + d_{1,4})/3 = (122 + 76 + 53)/3 = 83.67.$$

This gives the dendogram shown in Figure 1.

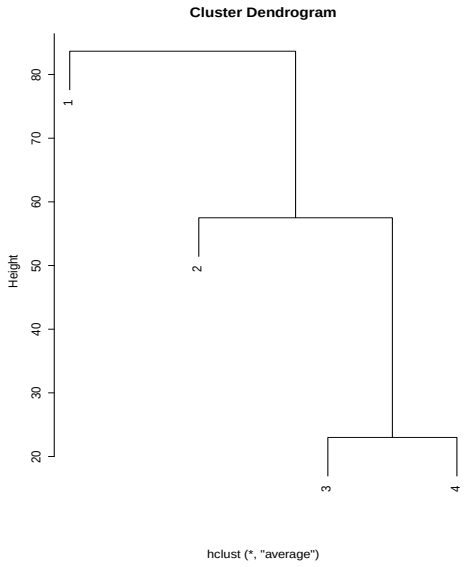


Figure 1: Dendogram for Question 4.

5. Suppose in an unsupervised learning application we are given observations $\dot{x}_1, \ldots, \dot{x}_n$. Recall the *within cluster sum of squares*, for K clusters $A_1, \ldots, A_K$ where d is a distance function and $g(A_i)$ is a cluster centroid:

$$SS_{within} = \sum_{i=1}^{K} \sum_{j \in A_i} d(\dot{x}_j, g(A_i))^2.$$

A K-means clustering algorithm was applied to the data, allowing the number of clusters K to vary from 1 to 4. The following table gives the separate sum of squares within each cluster:

	1	2	3	4
1	113.1	-	-	-
2	26.3	20.7	-	-
3	5.8	8.4	11.5	-
4	0.7	4.1	4.3	3.9

How can we estimate the proportion R^2 of total variation explained by the clusters? Sketch a plot of this proportion at a function of K. If we accept as the number of clusters the smallest value of K for which $R^2 \geq 80\%$, what is this number?

SOLUTION: The total sum of squares SS_{total} is simply the SS for the $K = 1$ model, so

$$SS_{total} = 113.1.$$

Otherwise, SS_{within} is the sum of the individual cluster sums of squares. Then

$$R^2 = 1 - \frac{SS_{within}}{SS_{total}}.$$

This gives, for $K = 1, 2, 3, 4$:

$$\begin{aligned} R^2[1] &= 1 - \frac{SS_{total}}{SS_{total}} = 0, \\ R^2[2] &= 1 - \frac{26.3 + 20.7}{113.1} = 0.584, \\ R^2[3] &= 1 - \frac{5.8 + 8.4 + 11.5}{113.1} = 0.773, \\ R^2[4] &= 1 - \frac{0.7 + 4.1 + 4.3 + 3.9}{113.1} = 0.885. \end{aligned}$$

The require plot is shown in Figure 2. The smallest number of clusters that yield at least 80% variation explained is $K = 4$.

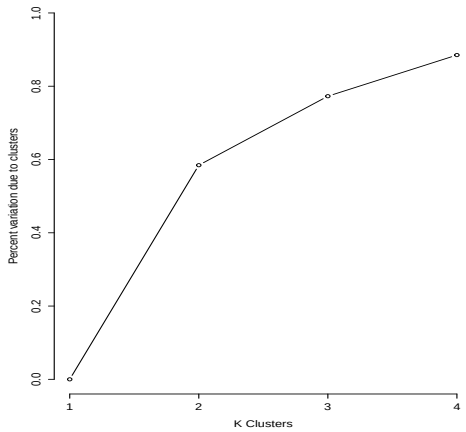


Figure 2: Plot for Question 5.

6. We observe n observations of the random vector $\tilde{X} = (X_1, X_2, X_3)$. Each component has zero mean and unit variance. The correlation matrix is

$$\Lambda = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}.$$

A scaled principal components analysis is performed on the data, yielding the following principle component loadings:

	PC1	PC2	PC3
X_1	0.73	-0.02	-0.68
X_2	0.68	-0.02	0.73
X_3	-0.03	-1.00	0.00

and the scree plot shown in Figure 3. Which of the correlations in Λ could plausibly be 0, and which are likely to be positive? Explain your answer.

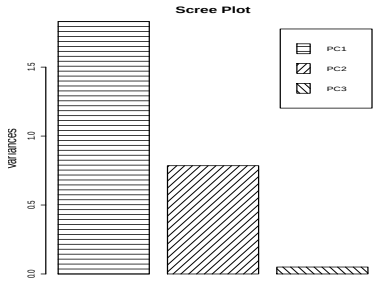


Figure 3: Scree plot for Question 6.

SOLUTION: By the scree plot, most variation is explained by the first 2 PCs. The first PC has loadings primarily on X_1, X_2, of the same sign. We therefore expect $\rho_{12} = \rho_{21} > 0$. Most of the loadings on the second PC is on X_3, so we expect X_3 to be approximately independent of X_1 and X_2, this means we could have $\rho_{13} = \rho_{31} = \rho_{23} = \rho_{32} = 0$.

7. Suppose we observe a geometric random variable $X \sim geom(p)$, which has probability mass function

$$P(k) = (1-p)^{k-1}p,$$

$k = 1, 2, \ldots$, for some $p \in (0, 1)$. Assume that the parameter p has a prior density $\pi(p)$:

$$p \sim beta(\alpha, \beta),$$

for some fixed $\alpha, \beta > 0$. Recall that the beta density $beta(\alpha, \beta)$ has the form

$$f(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \quad u \in [0, 1].$$

Show that the beta prior density is a conjugate density for p, that is, that the posterior density for p given X is also a beta density. Give this density precisely.

SOLUTION: Recall that to evaluate a posterior density of a parameter θ given data x it is often easiest to first express it as

$$\pi(\theta \mid x) = Kg(\theta)$$

where K is a constant that does not depend on θ, and then normalize $g(\theta)$. In this case we have

$$\pi(p \mid x) \propto f(x \mid p)\pi(p)$$

where $x \mid p \sim geom(p)$ and $p \sim beta(\alpha, \beta)$. This means

$$\begin{aligned} \pi(p \mid x) &= K(1-p)^{x-1}p \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= Kp^{\alpha+1-1}(1-p)^{\beta+x-2}, \end{aligned}$$

where K does not depend on p. We conclude that the posterior density of p given x is

$$p \mid x \sim beta(\alpha + 1, \beta + x - 1).$$

8. The following 6 variables are defined:

$$\begin{aligned} X_1 &= \textit{Attends Concert} \\ X_2 &= \textit{Exposure to Bacteria} \\ X_3 &= \textit{Genetic Predisposition} \\ X_4 &= \textit{Infection} \\ X_5 &= \textit{Resistance to Antibiotics} \\ X_6 &= \textit{Misses School} \end{aligned}$$

and are used in the following model:

$$\begin{aligned} X_2 &= X_1 + \epsilon_1 \\ X_4 &= X_2 + X_3 + \epsilon_2 \\ X_6 &= X_4 + X_5 + \epsilon_3. \end{aligned}$$

The terms $\epsilon_1, \epsilon_2, \epsilon_3$ are independently distributed errors.

(a) Suppose we want to model $\tilde{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$ as a Bayesian network. Sketch all DAGs which imply conditional independencies consistent with $\tilde{X}$.

(b) Suppose we wish to develop a predictive model for response X_6 using the remaining elements of $\tilde{X}$ as predictors (we assume they are all observable). Which elements should we use? Answer the same question for response X_4.

SOLUTION:

(a) The equations imply 5 edges: $1 \to 2$, $2 \to 4$, $3 \to 4$, $4 \to 6$ and $5 \to 6$. This graph is shown Figure 4 (left). Any equivalent graph must be obtained by reversing the direction of some combination of edges without adding or removing a v-structure. There are 2 v-structrues: $2 \to 4$, $3 \to 4$ and $4 \to 6$, $5 \to 6$. None of those edges can be changed. However, the direction of $1 \to 2$ can be changed, yielding the 2 equivalent DAGs in Figure 4.

(b) The *Markov blanket* of a node is that node's children, parents and children's other parents. A node is independent of the remaining nodes conditional on it's Markov blanket. The Markov blanket of X_6 is $\{X_4, X_5\}$, so we would only need to use X_4 and X_5 in a predictive model with response X_6. The Markov blanket of X_4 is $\{X_2, X_3, X_5, X_6\}$, so we would only need to use those variables in a predictive model with response X_4.

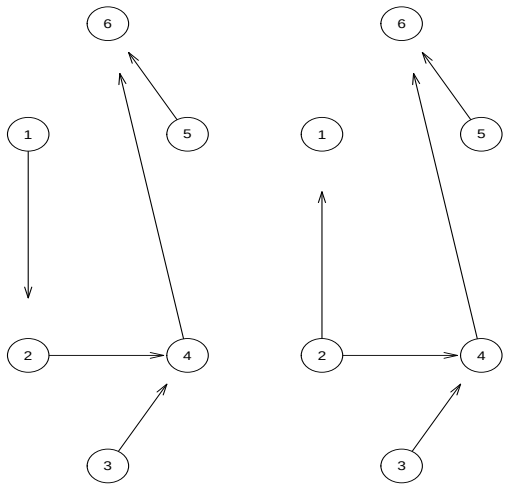


Figure 4: DAGs for Question 8(a).

9. The *Raleigh distribution* models positive random variables. It has one parameter $\sigma^2 > 0$ and a cumulative distribution function (CDF) given by

$$F(x) = 1 - e^{-x^2/(2\sigma^2)}, \quad x \geq 0.$$

Derive the survival function and the hazard function for the Raleigh distribution. Is a Raleigh survival time *new better than used* (NBU) or *new worse than used* (NWU)?

SOLUTION: We have survival function

$$S(x) = 1 - F(X) = e^{-x^2/(2\sigma^2)}, \quad x \geq 0.$$

The density $f(x)$ is the derivative of F:

$$f(x) = \frac{d1 - e^{-x^2/(2\sigma^2)}}{dx} = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)},$$

so that the hazard function is

$$h(x) = \frac{f(x)}{S(x)} = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} \div e^{-x^2/(2\sigma^2)} = \frac{x}{\sigma^2}, \quad x \geq 0.$$

The hazard rate is increasing with x so the survival time is *new better than used* (NBU).

10. Suppose we observe survival times $17, 20+, 20, 34$. Recall that the symbol '+' denotes a right-censored observation. Construct and sketch a Kaplan-Meier estimate for the survival function.

SOLUTION: For survival times $17, 20+, 20, 34$ we have table:

i	t_i	d_i	$r(t_i)$	$\hat{p}_i$
0	0	0	4	(4-0)/4 = 1
1	17	1	4	(4-1)/4 = 3/4
2	20	1	3	(3-1)/3 = 2/3
3	34	1	1	(1-1)/1 = 0

Then plot the cumulative products

$$\hat{p}_0, \hat{p}_0\hat{p}_1, \hat{p}_0\hat{p}_1\hat{p}_2, \hat{p}_0\hat{p}_1\hat{p}_2\hat{p}_3 = 1, 3/4, 1/2, 0$$

at times

$$t_0, \ldots, t_3 = 0, 17, 20, 34.$$

Note that '+' indicates the position of a censored observation. See Figure 5.

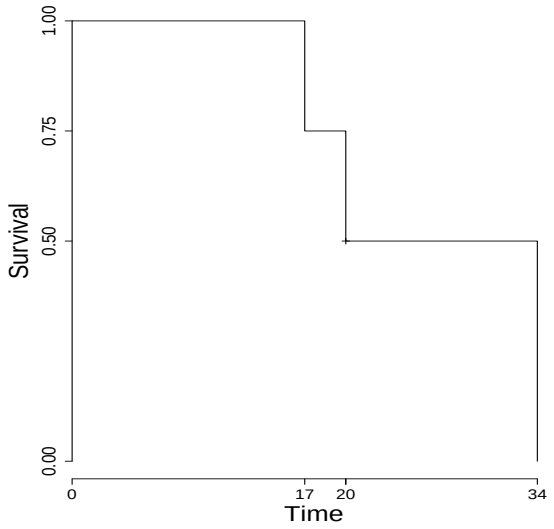


Figure 5: Kaplan-Meier estimate of survival function for Question 10

Final Exam - CSC/DSC 265/465 - May 12, 2017

NAME: _____

You are allowed up to 5 aid sheets on standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). You have 3 hours. Answer all eight questions. All questions have equal weight. You are encouraged to read each question completely before starting.

Q1: The following full linear regression model is considered:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

An all-subsets model selection procedure is to be used to determine which of the predictors and interactions to retain. The relevant SSE values are given in the following table. The sample size is $n = 25$. Which model possesses the largest coefficient of determination R^2? Which model possesses the largest adjusted coefficient of determination R^2_{adj}?

	Model	SSE
1	y=1	10638.06
2	y=x1	7810.40
3	y=x2	2210.16
4	y=x1+x2	29.23
5	y=x1+x2+x1*x2	27.39

SOLUTION: The model with the highest R^2 must be model 5, since all other models are reduced models.

The total sum of squares is given by the null model (model 1):

$$SSTO = 10638.06$$

The formula is

$$R^2_{adj} = 1 - \frac{SSE/(n-(q+1))}{SSTO/(n-1)}.$$

where q is the number of predictors. We can construct table:

	Model	SSE	q	R^2_{adj}
1	y=1	10638.06	0	0.00000
2	y=x1	7810.40	1	0.23388
3	y=x2	2210.16	1	0.78321
4	y=x1+x2	29.23	2	0.99700
5	y=x1+x2+x1*x2	27.39	3	0.99706

Model 5 has the highest R^2_{adj}.

Q2: We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and x_i is a predictor variable. Suppose $g(x)$ is a piecewise polynomial with knots at $\xi = 0, 1$ of the form:

$$g(x) = \begin{cases} \alpha & ; \quad x \leq 0 \\ ax^3 + bx^2 + cx + d & ; \quad x \in (0, 1] \\ \beta & ; \quad x > 1 \end{cases} .$$

It is further required that $g(x)$ is continuous at both knots, and that the first derivative of $g(x)$ is continuous and equal to 0 at both knots. See Figure 1 for an example of $g(x)$.

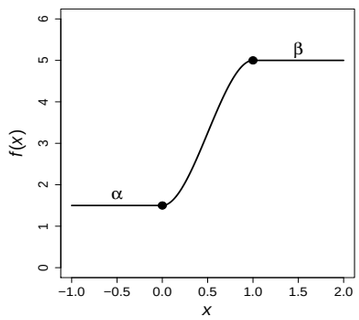


Figure 1: Example of $g(x)$ for Question **Q2**.

(a) Express the coefficients a, b, c, d as functions of α and β.

(b) How many degrees of freedom does the model have?

SOLUTION:

(a) **[7]** Note that

$$\frac{d}{dx}(ax^3 + bx^2 + cx + d) = 3ax^2 + 2bx + c.$$

So, we have constraints:

$$\begin{aligned} \alpha &= d \\ 0 &= c \\ \beta &= a + b + c + d \\ 0 &= 3a + 2b + c. \end{aligned}$$

Directly, we have $c = 0$, $d = \alpha$. Substitution gives

$$\begin{aligned} a + b &= \beta - \alpha \\ 3a + 2b &= 0. \end{aligned}$$

To summarize:

$$\begin{aligned} a &= 2(\alpha - \beta) \\ b &= -3(\alpha - \beta) \\ c &= 0 \\ d &= \alpha. \end{aligned}$$

(b) **[3]** 6 parameters with 4 constraints yields $2 = 6 - 4$ model degrees of freedom.

Q3: We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and x_i is a predictor variable. We consider the following five models

M1 $g(x) = \beta_0 + \beta_1 x$, where β_0, β_1 are to be estimated.

M2 $g(x) = \beta_1 / \sqrt{x}$, where β_1 is to be estimated.

M3 $g(x) = \beta_0 + \beta_1 / \sqrt{x}$, where β_0, β_1 are to be estimated.

M4 $g(x)$ is a continuous piecewise linear spline with 2 knots at $\xi = 3, 8$.

M5 $g(x)$ is a cubic spline with 1 knot at $\xi = 5$.

The relevant SSE values are given in the following table. The sample size is $n = 91$. Which model is preferred based on the AIC score (use form $AIC = n \log(SSE/n) + C$).

Model	SSE
M1	3.171
M2	2.814
M3	2.812
M4	2.801
M5	2.780

SOLUTION: The formula is

$$AIC = n \log(SSE/n) + 2k,$$

where k is the number of parameters. Other than σ^2, the number of parameters is

M1 β_0, β_1, $k = 2$.

M2 β_1, $k = 1$.

M3 β_0, β_1, $k = 2$.

M4 $2 + N_{knots}$, where $N_{knots} = 2$ is the number of knots, so $k = 4$.

M5 $4 + N_{knots}$, where $N_{knots} = 1$ is the number of knots, so $k = 5$.

We can construct table:

	Model	SSE	k	AIC	$k+1$	AIC (with σ^2)
1	M1	3.171	2	-301.481	3	-299.481
2	M2	2.814	1	-314.339	2	-312.339
3	M3	2.812	2	-312.400	3	-310.400
4	M4	2.801	4	-308.749	5	-306.749
5	M5	2.780	5	-307.452	6	-305.452

So, model **M2** has the lowest AIC (including σ^2 in the parameter count necessarily yields the same conclusion).

Q4: Suppose we have $n = 5$ observations of features, for which the following distance matrix is calculated:

	1	2	3	4	5
1	0.00	14.29	15.43	17.25	16.45
2	14.29	0.00	12.36	17.50	15.52
3	15.43	12.36	0.00	18.13	16.22
4	17.25	17.50	18.13	0.00	13.84
5	16.45	15.52	16.22	13.84	0.00

Using the single link agglomeration method, construct a hierarchical cluster for this data. Sketch a dendogram, indicating precisely the height of each node.

SOLUTION: The single link distance between two clusters A and B is

$$D(A, B) = \min_{i \in A, j \in B} d_{ij}.$$

To construct the clustering, we use the following steps:

1. Start with clusters {1},{2},{3},{4},{5}.
2. First join the two nearest observations, which are 2 and 3 ($d_{2,3} = 12.36$). This gives clusters $\{1\}$, $\{4\}$, $\{5\}$ and $\{2, 3\}$ joined at distance 12.36.
3. The cluster distances are now

$$\begin{aligned} D(\{1\}, \{4\}) &= d_{1,4} = 17.25, \\ D(\{1\}, \{5\}) &= d_{1,5} = 16.45, \\ D(\{4\}, \{5\}) &= d_{4,5} = 13.84, \\ D(\{1\}, \{2, 3\}) &= \min\{d_{1,2}, d_{1,3}\} = \min\{14.29, 15.43\} = 14.29, \\ D(\{4\}, \{2, 3\}) &= \min\{d_{4,2}, d_{4,3}\} = \min\{17.50, 18.13\} = 17.50, \\ D(\{5\}, \{2, 3\}) &= \min\{d_{5,2}, d_{5,3}\} = \min\{15.52, 16.22\} = 15.52. \end{aligned}$$

   The smallest cluster distance is $D(\{4\}, \{5\}) = 13.84$, so combine clusters $\{4\}$ and $\{5\}$. This gives clusters $\{1\}$, $\{2, 3\}$ and $\{4, 5\}$, joined at distance 13.84.
4. The cluster distances are now

$$\begin{aligned} D(\{1\}, \{2, 3\}) &= \min\{d_{1,2}, d_{1,3}\} = \min\{14.29, 15.43\} = 14.29, \\ D(\{1\}, \{4, 5\}) &= \min\{d_{1,4}, d_{1,5}\} = \min\{17.25, 16.45\} = 16.45, \\ D(\{2, 3\}, \{4, 5\}) &= \min\{d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\} = \min\{17.50, 15.52, 18.13, 16.22\} = 15.52. \end{aligned}$$

   The smallest cluster distance is $D(\{1\}, \{2, 3\}) = 14.29$, so combine clusters $\{1\}$ and $\{2, 3\}$. This gives clusters $\{1, 2, 3\}$ and $\{4, 5\}$, joined at distance 14.29.
5. Join the remaining two clusters, at cluster distance

$$D(\{1, 2, 3\}, \{4, 5\}) = \min\{d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\} = \{17.25, 16.45, 17.50, 15.52, 18.13, 16.22\} = 15.52.$$

This gives the dendogram shown in Figure 2.

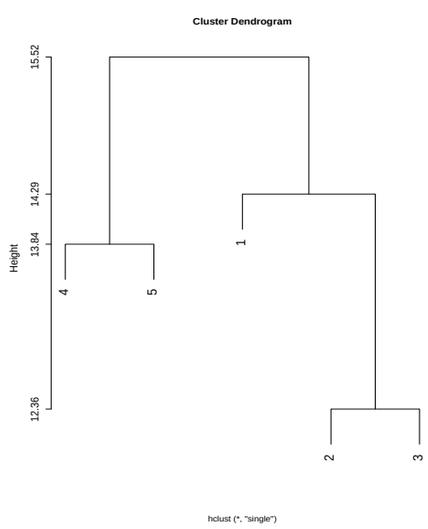


Figure 2: Dendogram for Question **Q4**.

Q5: Suppose in an unsupervised learning application we are given observations $\dot{x}_1, \ldots, \dot{x}_n$. Recall the *within cluster sum of squares*, for K clusters $A_1, \ldots, A_K$ where d is a distance function and $g(A_i)$ is a cluster centroid:

$$SS_{within} = \sum_{i=1}^{K} \sum_{j \in A_i} d(\dot{x}_j, g(A_i))^2.$$

A K-means clustering algorithm was applied to the data, allowing the number of clusters K to vary from 1 to 6. The following table gives the separate sum of squares within each cluster:

	1	2	3	4	5	6
1	12679.5	-	-	-	-	-
2	1741.7	1503.0	-	-	-	-
3	606.8	399.8	446.0	-	-	-
4	197.8	161.4	188.6	177.3	-	-
5	57.4	182.6	151.9	89.6	64.8	-
6	93.6	49.0	86.6	50.6	34.0	79.0

Let R^2 be the proportion of total variation explained by the clustering. If we accept as the number of clusters the smallest value of K for which $R^2 \geq 90\%$, what is this number?

SOLUTION: The total sum of squares SS_{total} is simply the SS for the $K = 1$ model, so

$$SS_{total} = 12679.5.$$

Otherwise, SS_{within} is the sum of the individual cluster sums of squares. Then

$$R^2 = 1 - \frac{SS_{within}}{SS_{total}}.$$

This gives, for $K = 1, 2, 3, 4$:

$$\begin{aligned}
R^2[1] &= 1 - \frac{SS_{total}}{SS_{total}} = 0, \\
R^2[2] &= 1 - \frac{1741.7 + 1503.0}{12679.5} = 0.7441, \\
R^2[3] &= 1 - \frac{606.8 + 399.8 + 446.0}{12679.5} = 0.8854, \\
R^2[4] &= 1 - \frac{197.8 + 161.4 + 188.6 + 177.3}{12679.5} = 0.9428. \\
R^2[5] &= 1 - \frac{57.4 + 182.6 + 151.9 + 89.6 + 64.8}{12679.5} = 0.9569. \\
R^2[6] &= +1 - \frac{93.6 + 49.0 + 86.6 + 50.6 + 34.0 + 79.0}{12679.5} = 0.9690.
\end{aligned}$$

The smallest number of clusters that yield at least 90% variation explained is $K = 4$.

Q6: Suppose for a certain application, data takes the form (X_1, X_2), where X_i are independent Bernoulli random variables (ie. assume value 0 or 1) of mean p. The usual unbiased estimator of p is

$$\hat{p}_1 = \frac{X_1 + X_2}{2},$$

which has variance $\sigma_p^2 = p(1-p)/2$. Suppose we also consider as an alternative estimator of p:

$$\hat{p}_2 = \hat{p}_1/2 + 1/4.$$

(a) Give an expression for $MSE = variance + bias^2$ as a function of p for each estimator.

(b) Which estimator has smallest MSE when $p = 1/2$?

(c) Which estimator has smallest MSE when $p = 1/8$?

SOLUTION:

(a) **[7]** For $\hat{p}_1$ we have:

$$\begin{aligned} var(\hat{p}_1) &= \frac{1}{2} \cdot p(1-p) \\ bias(\hat{p}_1) &= E[\hat{p}_1] - p = p - p = 0 \\ MSE(\hat{p}_1) &= var(\hat{p}_1) + bias(\hat{p}_1)^2 = \frac{1}{2} \cdot p(1-p). \end{aligned}$$

For $\hat{p}_2$ we have:

$$\begin{aligned} var(\hat{p}_2) &= var(\hat{p}_1/2 + 1/4) = \frac{1}{2^2} \cdot var(\hat{p}_1) = \frac{1}{8} \cdot p(1-p) \\ bias(\hat{p}_2) &= E[\hat{p}_2] - p = p/2 + 1/4 - p = -p/2 + 1/4 \\ MSE(\hat{p}_1) &= var(\hat{p}_2) + bias(\hat{p}_2)^2 = \frac{1}{8} \cdot p(1-p) + (p/2 - 1/4)^2 = \frac{1}{8} \cdot (p^2 - p + 1/2). \end{aligned}$$

(b) **[1.5]** For $p = 1/2$,

$$\begin{aligned} MSE(\hat{p}_1) &= \frac{1}{2} \cdot 1/2(1 - 1/2) = 1/8, \\ MSE(\hat{p}_2) &= \frac{1}{8} \cdot ((1/2)^2 - 1/2 + 1/2) = 1/32. \end{aligned}$$

so $MSE(\hat{p}_2) < MSE(\hat{p}_1)$.

(c) **[1.5]** For $p = 1/8$,

$$\begin{aligned} MSE(\hat{p}_1) &= \frac{1}{2} \cdot 1/8(1 - 1/8) = 7/128 = 7/2^7, \\ MSE(\hat{p}_2) &= \frac{1}{8} \cdot ((1/8)^2 - 1/8 + 1/2) = 25/512 = 25/2^9. \end{aligned}$$

so $0.0488 = MSE(\hat{p}_2) < MSE(\hat{p}_1) = 0.0547$.

Q7: There is often interest in determining whether or not two quantities X and Y have a *power-law* relationship:

$$Y = aX^b \tag{1}$$

for two constants a, b. Suppose, given $n = 41$ independent paired observations (X_i, Y_i) of these quantities, we fit model

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \beta_2 \log(X_i)^2 + \epsilon_i, \;\; i = 1, \ldots, n,$$

assuming any relevant distributional assumption holds, and get output:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.33180   19.50917   1.965   0.0568 .
log.x         -5.93521    2.67072  -2.222   0.0323 *
log.x.squared  0.02227    0.08868   0.251   0.8031
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) How can the output be used to assess the validity of the power-law relationship of Equation (1)? What do you conclude?

(b) Give level 95% confidence intervals for parameters a and b, based on this output.

SOLUTION:

(a) **[5]** If we take a log-transform of the model $\log(Y) = \log(a) + b\log(X)$. We can therefore equate $\beta_0 = \log(a)$ and $\beta_1 = b$ in the linear model. In addition, if the model holds we must have $\beta_2 = 0$. The 2-sided p-value for rejecting null hypothesis $H_o : \beta_2 = 0$ is $p = 0.8031$ from the output. We do not reject the power-law relationship, therefore.

(b) **[5]** Since $\beta_0 = \log(a)$ and a 95% CI for β_0 is

$$\hat{\beta}_0 \pm 2 \times SE = 38.33180 \pm 2 \times 19.50917 = (-0.68654, 77.35014)$$

a 95% CI for a is

$$e^{\hat{\beta}_0 \pm 2 \times SE} = e^{38.33180 \pm 2 \times 19.50917} = (0.5033, 3.9 \times 10^{33}).$$

Since $\beta_1 = b$ a 95% CI for b is identical to the 95% CI for β_1:

$$\hat{\beta}_0 \pm 2 \times SE = -5.93521 \pm 2 \times 2.67072 = (-11.27665, -0.59377).$$

Q8: Data for a linear predictive model consists of responses Y, and five feature variables $X_1, \ldots, X_5$. There are $n = 100$ observations. Each feature is standardized to a mean of 0 and a standard deviation of 1. The principal components $PC_1, \ldots, PC_5$ are calculated. The coefficients (loadings) are given below:

```
        PC1         PC2         PC3          PC4         PC5
X1 0.4846410  0.02535008 -0.64416209  0.4756905738 -0.35107564
X2 0.4967931  0.04064938 -0.05731951 -0.8099920823 -0.30359769
X3 0.4659406  0.24895000  0.74092022  0.3366118674 -0.24218360
X4 0.5001321  0.13595192 -0.08208782  0.0003119937  0.85126166
X5 0.2260240 -0.95772959  0.16145740  0.0657545280  0.03570757
```

The following linear regression models (along with error sum of squares SSE) are fit:

Model	SSE
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$	2287.042
$Y = \beta_0 + \beta_1 PC_1 + \epsilon$	10264.414
$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \epsilon$	2150.351
$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \epsilon$	2150.351

(a) Which model is preferred based on the BIC score (use form $BIC = n \log(SSE/n) + C$)?

(b) Suppose the predictor is to take the form of a linear combination of the five feature variables, plus an intercept term. One of the three strategies is to be used:

(i) Use whichever linear combination minimizes the SSE.

(ii) Use only the unweighted mean $\bar{X} = (X_1 + \ldots + X_5)/5$ of the feature variables, that is $Y \approx \beta_0 + \beta_1 \bar{X}$.

(iii) Use the predictor of part (ii) but add one more feature, say, $m \in \{1, 2, 3, 4, 5\}$, that is $Y \approx \beta_0 + \beta_1 \bar{X} + \beta_2 X_m$.

Based on your BIC analysis, and the principal component loadings, which of the three strategies seems preferable?

SOLUTION:

(a) **[7]** The formula is

$$BIC = n \log(SSE/n) + log(n)k,$$

where k is the number of parameters. We can construct table:

	Model	SSE	k	BIC
1	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$	2287.042	6	340.615
2	$Y = \beta_0 + \beta_1 PC_1 + \epsilon$	10264.414	2	472.337
3	$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \epsilon$	2150.351	3	320.637
4	$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \epsilon$	2150.351	4	325.242

Model 3 $Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \epsilon$ is selected by the BIC score.

(b) **[3]** For PC_1 the loadings are nearly equal for X_1, X_2, X_3, X_4. On the other hand, PC_2 is dominated by X_5. The best BIC model is based on PC_1, PC_2, and most closely resembles strategy (iii) with $X_m = X_5$.

Final Exam Practice Questions - Set 1 - CSC 265 - Spring 2016

1. A Bayesian network model is fit for 8 genes labeled $a - h$, resulting in the following DAG. We say a gene y is downstream from gene x if there is a directed path from x to y.

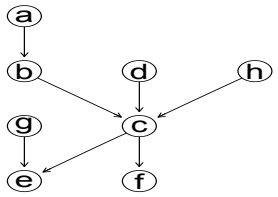


Figure 1: Sample DAG for Question 1

   (a) List all v-structures of the DAG.

   (b) State precise conditions under which two DAGs are equivalent (that is, are members of the same equivalent class).

   (c) Suppose the DAG is accepted as a true model of regulatory control. Then, in reference to the equivalence classes of the DAG, a statement about regulatory order may be one of three types:

      A. Implied by the Bayesian network model (true of all equivalent DAGs).
      B. Compatible with the Bayesian network model (true of some but not all equivalent DAGs).
      C. Not compatible with the Bayesian network model (not true of any equivalent DAG).

   Note that a DAG is equivalent to itself. Of which type (A, B or C) is each of the following statements? Briefly justify your answer.

      i. c is downstream from h.
      ii. g is downstream from c.
      iii. h has no parents.
      iv. b has no parents.
      v. c has exactly three parents.
      vi. f is downstream from a.

2. Suppose we observe a normally distributed random variable $X \sim N(\mu, \sigma)$. Assume σ is known, and that μ has a prior density $\pi(\mu)$:

$$\mu \sim N(\mu_0, \sigma_0),$$

   for some fixed μ_0, σ_0. Show that the normal prior density is a conjugate density for μ, that is, that the posterior density for μ given X is also normal. Give this density precisely.

3. The *gamma density*, denoted $X \sim gamma(\alpha, \beta)$ is defined as

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

   where $x, \alpha, \beta > 0$. Suppose we observe a Poisson random variable $X \sim pois(\lambda)$, and that λ has a prior density $\pi(\lambda)$:

$$\lambda \sim gamma(\alpha_0, \beta_0)$$

   for some fixed α_0, β_0. Show that the gamma prior density is a conjugate density for λ, that is, that the posterior density for λ given X is also gamma. Give this density precisely.

4. Suppose we observe survival times $45, 55+, 80, 90$. Construct and sketch a Kaplan-Meier estimate for the survival function.

Final Exam Practice Questions - Set 1 - CSC 265 - Spring 2016 - SOLUTIONS

1. (a) There 4 v-structures:
   i. $b \to c$, $d \to c$
   ii. $b \to c$, $h \to c$
   iii. $d \to c$, $h \to c$
   iv. $g \to e$, $c \to e$

   (b) v-structures and topologies are identical.

   (c) i. A. $h \to c$ is part of (at least) one v-structure which cannot be removed.
   ii. C. $g \to e$ and $c \to e$ form a v-structure, which cannot be removed.
   iii. A. h shares an edge with c only, which is part of a v-structure and cannot be changed.
   iv. B. The direction of the edge $a \to b$ can be changed to yield an equivalent graph in which b has no parents.
   v. A. All parents of c are part of v-structures pointing to c. Deletion or addition of any other parent would add or delete a v-structure.
   vi. B. The direction of the edge $a \to b$ can be changed to yield an equivalent graph in which f is not downstream of a.

2. Recall that to evaluate a posterior density of a parameter θ given data x it is often easiest to first express it as

$$\pi(\theta \mid x) = Kg(\theta)$$

where K is a constant that does not depend on θ, and then normalize $g(\theta)$. This means we don't need to actually evaluate K. In this case we have

$$\pi(\mu \mid x) \propto f(x \mid \mu)\pi(\mu)$$

where $x \mid \mu \sim N(\mu, \sigma)$ and $\mu \sim N(\mu_0, \sigma_0)$. This means

$$\pi(\mu \mid x) = Ke^{-\frac{1}{2}Q_1}e^{-\frac{1}{2}Q_0}$$

where

$$Q_1 = \frac{(x-\mu)^2}{\sigma^2}, \;\; Q_0 = \frac{(\mu-\mu_0)^2}{\sigma_0^2}.$$

We then have

$$Q_1 + Q_0 = \mu^2\left[\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right] - 2\mu\left[\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right] + \left[\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right].$$

This means

$$\pi(\mu \mid x) = Ke^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right] - 2\mu\left[\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right]\right\}}$$

where K does not depend on μ, and that $\pi(\theta \mid x) \sim N(\mu_{post}, \sigma^2_{post})$ is a normal density function with mean and variance

$$\begin{aligned} \mu_{post} &= \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}, \\ \sigma^2_{post} &= \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}. \end{aligned}$$

3. Recall that to evaluate a posterior density of a parameter θ given data x it is often easiest to first express it as

$$\pi(\theta \mid x) = Kg(\theta)$$

where K is a constant that does not depend on θ, and then normalize $g(\theta)$. This means we don't need to actually evaluate K. In this case we have

$$\pi(\lambda \mid x) \propto f(x \mid \lambda)\pi(\lambda)$$

where $x \mid \lambda \sim pois(\lambda)$ and $\lambda \sim gamma(\alpha_0, \beta_0)$. This means

$$\begin{aligned} \pi(\lambda \mid x) &= K\lambda^x e^{-\lambda}\lambda^{\alpha_0-1}e^{-\beta_0\lambda} \\ &= K\lambda^{x+\alpha_0-1}e^{-(\beta_0+1)\lambda}, \end{aligned}$$

where K does not depend on λ. We conclude that the posterior density of λ given x is

$$\lambda \mid x \sim gamma(x + \alpha_0, \beta_0 + 1).$$

4. For survival times $45, 55+, 80, 90$ we have table:

i	t_i	d_i	$r(t_i)$	$\hat{p}_i$
0	0	0	4	(4-0)/4 = 1
1	45	1	4	(4-1)/4 = 3/4
2	55	0	3	(3-0)/3 = 1
3	80	1	2	(2-1)/2 = 1/2
4	90	1	1	(1-1)/1 = 0

Then plot the cumulative products $\hat{p}_0, \hat{p}_0\hat{p}_1, \ldots, \hat{p}_0\hat{p}_1\hat{p}_2\hat{p}_3\hat{p}_4$ at times $t_0, \ldots, t_4$. Note that '+' indicates the position of a censored observation. See Figure 1.

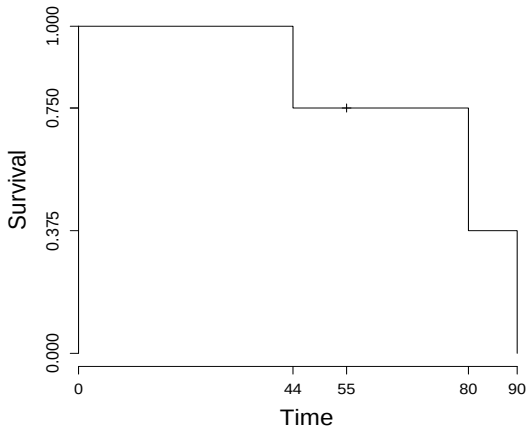


Figure 1: Kaplan-Meier estimate of survival function for Question 4

Final Exam Practice Questions - Set 2 - CSC 265 - Spring 2016

1. Two variables Y and X are believed to have the following relationship:

$$Y = aX^b$$

for two constants a, b. There is special interest in knowing whether or not this relationship is concave (equivalently, $b < 1$). In order to resolve this question paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled, where $n = 34$. The simple linear regression model

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

is fit, with the following output:

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.7115358 0.09647197  7.37557 2.174990e-08
log(x)      0.8948970 0.03583328 24.97391 1.483693e-22
```

Formulate appropriate null and alternative hypotheses for this question in terms of the regression coefficients β_0 and/or β_1. Is there evidence at an $\alpha = 0.05$ significance level that $b < 1$?

2. Given a single predictor x and response y, a polynomial regression model is considered:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon.$$

Suppose we consider four models (full model and 3 reduced models). Suppose further that the sample size is $n = 75$, and that the four models are fit, yielding the following error sums of squares SSE:

Table 1:

	MODEL	SSE
M_0	$y = \beta_0 + \epsilon$	239.2
M_1	$y = \beta_0 + \beta_1 x + \epsilon$	82.8
M_2	$y = \beta_0 + \beta_2 x^2 + \epsilon$	87.9
M_{12}	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$	79.8

Determine the model selected by the AIC and the BIC criterion. Use the form $n \log(SSE/n) + C$.

3. Suppose we have $n = 4$ observations of features, for which the following distance matrix is calculated:

	1	2	3	4
1	0.00	0.18	0.37	0.30
2	0.18	0.00	0.19	0.13
3	0.37	0.19	0.00	0.06
4	0.30	0.13	0.06	0.00

Using the maximum distance (ie. complete, or compact) agglomeration method, construct a hierarchical cluster for this data. Sketch a dendogram, indicating precisely the height of each node.

4. Suppose in an unsupervised learning application we are given observations $\dot{x}_1, \ldots, \dot{x}_n$. Recall the *within cluster sum of squares*, for K clusters $A_1, \ldots, A_K$ where d is a distance function and $g(A_i)$ is a cluster centroid:

$$SS_{within} = \sum_{i=1}^{K} \sum_{j \in A_i} d(\dot{x}_j, g(A_i))^2.$$

A K-means clustering algorithm was applied to the data, allowing the number of clusters K to vary from 1 to 4. The following table gives the separate sum of squares within each cluster:

K	1	2	3	4
1	101.5	-	-	-
2	13.8	24.1	-	-
3	5.5	4.0	8.2	-
4	5.5	2.2	0.4	2.9

How can we estimate the proportion R^2 of total variation explained by the clusters? Sketch a plot of this proportion at a function of K. If we accept as the number of clusters the smallest value of K for which $R^2 \geq 80\%$, what is this number?

Final Exam Practice Questions - Set 2 - CSC 265 - Spring 2016 - SOLUTIONS

1. The appropriate hypotheses are $H_o : b \geq 1$ and $H_a : b < 1$, since we are looking for evidence that $b < 1$. In terms of the regression coefficients this is equivalent to

$$H_o : \beta_1 \geq 1 \text{ and } H_a : \beta_1 < 1.$$

The test statistic is

$$T = \frac{\hat{\beta}_1 - 1}{S_{\hat{\beta}_1}} = \frac{0.8948970 - 1}{0.03583328} = -2.933112,$$

which has a t-distribution with $n - 2$ degrees of freedom under H_o. Since $t_{32,0.05} = 1.69$, we reject H_o at a 0.05 significance level, and conclude that $b < 1$.

2. We have scores:

$$\begin{aligned} AIC &= n \log(SSE/n) + 2q, \\ BIC &= n \log(SSE/n) + \log(n)q, \end{aligned}$$

where q is the number of parameters, and $n = 75$. This leads to table:

Model	q	AIC	BIC
M_0	1	88.99	91.30
M_1	2	11.42	16.06
M_2	2	15.90	20.54
M_3	3	10.65	17.61

The AIC selection is M_3, while the BIC selection is M_1.

3. The compact distance between two clusters A and B is

$$D(A, B) = \max\{d_{ij} : i \in A, j \in B\}.$$

To construct the clustering, we use the following steps:

(a) Start with clusters {1},{2},{3},{4}.

(b) First join the two nearest observations, which are 3 and 4 ($d_{3,4} = 0.06$). This gives clusters {1}, {2},{3, 4}, joined at distance 0.06.

(c) The cluster distances are now

$$\begin{aligned} D(\{1\}, \{2\}) &= d_{1,2} = 0.18, \\ D(\{1\}, \{3, 4\}) &= \max\{d_{1,3}, d_{1,4}\} = \max\{0.30, 0.37\} = 0.37, \\ D(\{2\}, \{3, 4\}) &= \max\{d_{2,3}, d_{2,4}\} = \max\{0.19, 0.13\} = 0.19. \end{aligned}$$

The smallest cluster distance is $D(\{1\}, \{2\}) = 0.18$, so combine clusters {1} and {2}. This gives clusters {1, 2},{3, 4}, joined at distance 0.18.

(d) The remaining clusters are joined. They have cluster distance

$$D(\{1, 2\}, \{3, 4\}) = \max\{d_{1,3}, d_{1,4}, d_{2,3}, d_{2,4}\} = \max\{0.37, 0.30, 0.19, 0.13\} = 0.37.$$

This gives the dendogram shown in Figure 1.

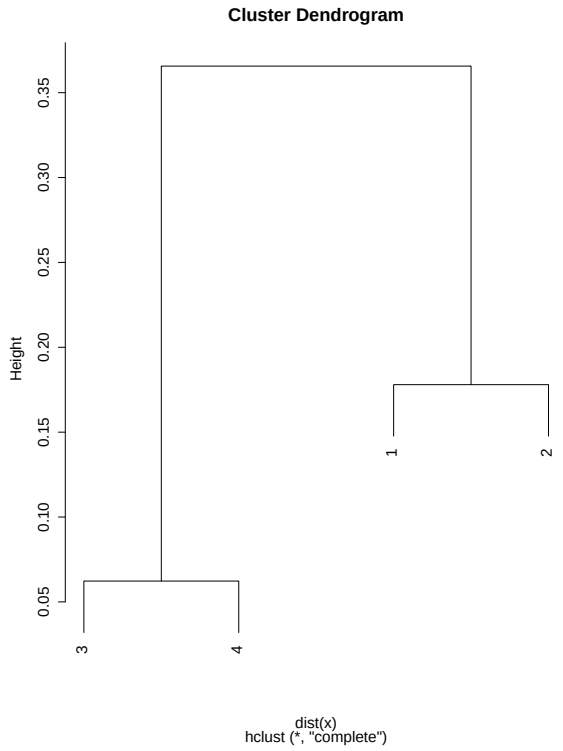


Figure 1: Dendogram for Question 3.

4. We are given cluster sums of squares:

K	1	2	3	4
1	101.5	-	-	-
2	13.8	24.1	-	-
3	5.5	4.0	8.2	-
4	5.5	2.2	0.4	2.9

The total sum of squares SS_{total} is simply the SS for the $K = 1$ model, so

$$SS_{total} = 101.05.$$

Otherwise, SS_{within} is the sum of the individual cluster sums of squares. Then

$$R^2 = 1 - \frac{SS_{within}}{SS_{total}}.$$

This gives, for $K = 1, 2, 3, 4$:

$$\begin{aligned} R^2[1] &= 1 - \frac{SS_{total}}{SS_{total}} = 0, \\ R^2[2] &= 1 - \frac{13.8 + 24.1}{101.5} = 0.627, \\ R^2[3] &= 1 - \frac{5.5 + 4.0 + 8.2}{101.5} = 0.826, \\ R^2[4] &= 1 - \frac{5.5 + 2.2 + 0.4 + 2.9}{101.5} = 0.892. \end{aligned}$$

The require plot is shown in Figure 2. The smallest number of clusters that yield at least 80% variation explained is $K = 3$.

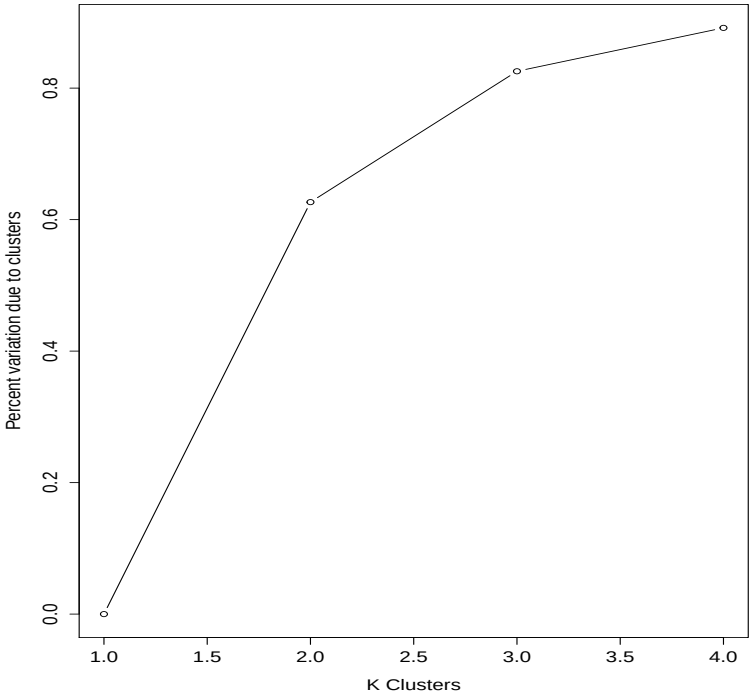


Figure 2: Plot for Question 4.

Final Exam Practice Questions - Set 3 - CSC 265 - Spring 2016

1. A regression model is to be developed for a response Y and single predictor X in range $X \in [10, 24]$, based on $n = 29$ paired observations. The following four models were considered, and the resulting error sum of squares is reported:

   (a) Constant model $Y = \beta_0 + \epsilon$ [SSE = 576983.05].

   (b) Simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ [SSE = 16546.78].

   (c) Linear spline with knots at $X = 15, 20$ [SSE = 13819.92].

   (d) Linear spline with knots at $X = 12, 17, 19, 22$ [SSE = 12169.24].

   Using the AIC score, which is the best model?

2. The *Pareto distribution* is sometimes used to model continuous survival times. It has two parameters $\theta_{min} > 0$ and $\alpha > 0$, and the density is given by

$$f(x \mid \theta_{min}, \alpha) = \begin{cases} \frac{\alpha(\theta_{min})^\alpha}{x^{\alpha+1}} & ; \quad x \geq \theta_{min} \\ 0 & ; \quad x < \theta_{min} \end{cases} .$$

   Derive the survival function and the hazard function for the Pareto distribution. Is a Pareto survival time *new better than used* (NBU) or *new worse than used* (NWU)?

3. A principal components analysis was performed on 4 measured psychometric scales: *Happiness*, *Joy*, *Mirth* and *Contentment*. The loadings on the PCs are given in the folowing table:

	PC1	PC2	PC3	PC4
Happiness	-0.58	0.07	-0.56	-0.58
Joy	-0.57	0.01	-0.23	0.79
Mirth	-0.58	0.01	0.79	-0.19
Contentment	0.05	1.00	0.03	0.03

   The scree plot for the principal components is shown in Figure 1. What conclusion can be reached from the loadings and the scree plot regarding the issue of dimension reduction?

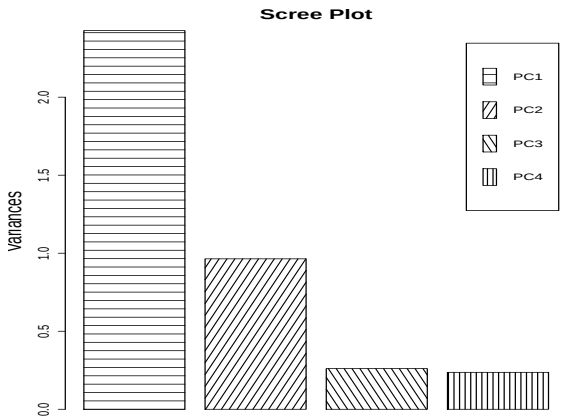


Figure 1: Scree plot for Question 3.

4. Suppose the density of a random vector $\tilde{X} = (X_1, X_2, X_3, X_4, X_5)$ can be factorized in the following way:

$$f(x_1, \ldots, x_5) = f(x_5 \mid x_4) f(x_4 \mid x_2, x_3) f(x_2 \mid x_1) f(x_1) f(x_3)$$

   (a) Suppose we want to model $\tilde{X}$ as a Bayesian network. Sketch all DAGs which imply conditional independencies consistent with $\tilde{X}$.

   (b) Suppose we wish to develop a predictive model for response X_5 using the remaining elements of $\tilde{X}$ as predictors. Which elements should we use? Answer the same question for response X_4.

Final Exam Practice Questions - Set 3 - CSC 265 - Spring 2016 - SOLUTIONS

1. We have two methods for calculating AIC:

$$\begin{aligned} AIC_1 &= \frac{1}{\hat{\sigma}^2} SSE + 2k \\ AIC_2 &= n \log(SSE/n) + 2k, \end{aligned}$$

where k is the number of parameters, and $\hat{\sigma}^2$ is an estimate of the error variance (here, use the estimate obtained from model 4, $\hat{\sigma}^2 = SSE/(n-6)$). For a linear spline with N knots and an intercept, we have $k = 2 + N$. This means the number of parameters for the four models is $k = 1, 2, 4, 6$, respectively. The AIC values are in the following table:

Model	k	AIC_1	AIC_2
1	1	1092.50	289.05
2	2	35.27	188.05
3	4	34.12	186.83
4	6	35.00	187.14

For both versions of the AIC, model 3 is the selected model.

2. Integrating the density gives

$$F(x) = \int_{u=\theta_{min}}^{x} f(u \mid \theta_{min}, \alpha) du = 1 - (\theta_{min}/x)^{\alpha}$$

for $x \geq \theta_{min}$. This gives survival and hazard functions

$$\begin{aligned} S(x) &= 1 - F(x) = (\theta_{min}/x)^{\alpha}, \\ h(x) &= f(x \mid \theta_{min}, \alpha)/S(x) = \alpha/x, \end{aligned}$$

for $x \geq \theta_{min}$. Since $h(x)$ is decreasing for all parameters, a Pareto survival time is always NWU.

3. From the scree plot, most of the variance is explained by the first two principal components. The loadings on the first PC are nearly equal for *Happiness*, *Joy* and *Mirth*, and relatively smaller for *Contentment*. The loadings on the second PC are concentrated on *Contentment*. We conclude that *Happiness*, *Joy* and *Mirth* are highly correlated, and together form a single dimension, while *Contentment* forms a second independent dimension.

4. (a) The factorization implies 4 edges: $4 \to 5$, $2 \to 4$, $3 \to 4$ and $1 \to 2$. This graph is shown Figure 1 (left). Any equivalent graph must be obtained by reversing the direction of some combination of edges without adding or removing a v-structure. Note that edges $2 \to 4$ and $3 \to 4$ form a v-structure, and so cannot be changed. If edge $4 \to 5$ is changed, then a new v-structure will be added. However, the direction of $1 \to 2$ can be changed, yielding the 2 equivalent DAGs in Figure 1.

   (b) The *Markov blanket* of a node is that node's children, parents and children's other parents. A node is independent of the remaining nodes conditional on it's Markov blanket. The Markov blanket of X_5 is X_4, so we would only need to use X_4 in a predictive model with response X_5. The Markov blanket of X_4 is $\{X_2, X_3, X_5\}$, so we would only need to use X_2, X_3, X_5 in a predictive model with response X_4.

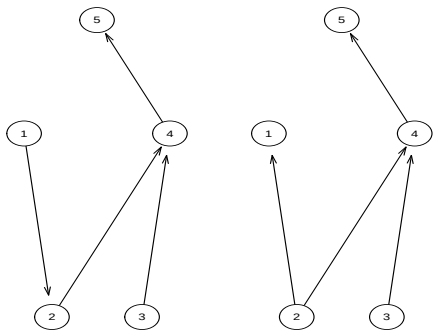


Figure 1: DAGs for Question 4(a).